

RESEARCH ARTICLE

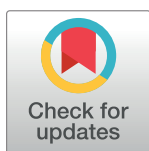
# Predicting seasonal influenza transmission using functional regression models with temporal dependence

Manuel Oviedo de la Fuente<sup>1,2\*</sup>, Manuel Febrero-Bande<sup>1,2</sup>, María Pilar Muñoz<sup>3,4</sup>, Ángela Domínguez<sup>4,5</sup>

**1** Technological Institute for Industrial Mathematics (ITMATI), Campus Vida, Santiago de Compostela, Spain, **2** MODESTYA Group, Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, Campus Vida, Santiago de Compostela, Spain, **3** Department of Statistics and Operation Research, Universitat Politècnica de Catalunya, Barcelona, Spain, **4** CIBER en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain, **5** Department of Medicine, Universitat de Barcelona, Barcelona, Spain

✉ Current address: Edif. Instituto Investigaciones Tecnológicas, planta -1, Rúa de Constantino Candeira s/n, 15782 Campus Vida, Santiago de Compostela, Spain

\* [manuel.oviedo@usc.es](mailto:manuel.oviedo@usc.es)



## OPEN ACCESS

**Citation:** Oviedo de la Fuente M, Febrero-Bande M, Muñoz MP, Domínguez À (2018) Predicting seasonal influenza transmission using functional regression models with temporal dependence. PLoS ONE 13(4): e0194250. <https://doi.org/10.1371/journal.pone.0194250>

**Editor:** Antonio Palazón-Bru, Universidad Miguel Hernandez de Elche, SPAIN

**Received:** July 6, 2017

**Accepted:** February 27, 2018

**Published:** April 25, 2018

**Copyright:** © 2018 Oviedo de la Fuente et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** This study was funded by the Catalan Agency for the Management of Grants for University Research (AGAUR Grant number 2014/SGR 1403) and cofunded by the Spanish Ministry of Economy and Competitiveness (Grant numbers MTM2013-41383-P and MTM2016-76969-P) and European Regional Development Fund (ERDF).

## Abstract

This paper proposes a novel approach that uses meteorological information to predict the incidence of influenza in Galicia (Spain). It extends the Generalized Least Squares (GLS) methods in the multivariate framework to functional regression models with dependent errors. These kinds of models are useful when the recent history of the incidence of influenza are readily unavailable (for instance, by delays on the communication with health informants) and the prediction must be constructed by correcting the temporal dependence of the residuals and using more accessible variables. A simulation study shows that the GLS estimators render better estimations of the parameters associated with the regression model than they do with the classical models. They obtain extremely good results from the predictive point of view and are competitive with the classical time series approach for the incidence of influenza. An iterative version of the GLS estimator (called iGLS) was also proposed that can help to model complicated dependence structures. For constructing the model, the distance correlation measure  $\mathcal{R}$  was employed to select relevant information to predict influenza rate mixing multivariate and functional variables. These kinds of models are extremely useful to health managers in allocating resources in advance to manage influenza epidemics.

## Introduction

Influenza is an infectious disease with person-to-person transmission that characteristically occurs as an epidemic affecting the whole population [1]. The influenza virus has been categorized into types A, B and C. However influenza C is a mild disease without seasonality and is therefore not considered in influenza epidemics. One remarkable feature of the influenza A

**Competing interests:** The authors have declared that no competing interests exist.

and B viruses is the frequency of changes in antigenicity. Alterations in the antigenic structure of the virus leads to infection by variants to which the population has little or no immunity.

The epidemiology of inter-pandemic influenza (also named seasonal influenza) is characterized in temperate zones by epidemics of variable size that occur during the colder winter months (November to April in the Northern Hemisphere and May to September in the Southern Hemisphere), each of which typically lasts 8–10 weeks [2]. In a study on influenza activity throughout eight seasons (1999–2007), the average length of epidemics in 23 European countries was 15.6 weeks (median 15 weeks; range 12–19 weeks) [3].

The reasons for the seasonal presentation of influenza epidemics are not entirely clear but they might result from more favourable environmental conditions for virus survival [4]. Various theories including improved virus survival in low temperatures, low humidity and low levels of ultraviolet radiation [2] have been advanced to explain this pattern in temperate zones. The typical incubation period for influenza is 1–4 days (average: 2 days).

Surveillance systems require accurate indicators that detect possible epidemics in advance. The epidemic of influenza is one of the problems of most concern to public health professionals across the world, due to its high levels of mortality and morbidity. Influenza is highly contagious and causes more morbidity than any other vaccine-preventable illness [5]. So, accurate estimates of the incidence of influenza are essential, for both public health services and citizens, to provide advance warning of epidemics and allow preventive measures to reduce contagion.

Statistical methods to forecast the incidence of influenza in particular, and contagious diseases in general, have changed over time. In one of the first studies on time series, Choi and Thacker [6] employed an ARIMA model to estimate pneumonia and influenza mortality. Dushoff et al. [7] used a regression model to investigate how cold temperatures contribute to excess seasonal mortality. Hohle and Paul [8] proposed an alternative model to monitor infectious diseases that consisted in applying count data charts to monitor time series. From a Bayesian framework, Conesa et al. [9] an automated monitoring of influenza surveillance data that made it possible to take the geographical component into account in statistical models in addition to temporal evolution was proposed. Contributions to this methodology are growing steadily through disease mapping. The studies by Ugarte et al. [10] and Paul and Held [11] are recent examples of this. Their common denominator is that they apply different statistical methodologies to multivariate time series (hierarchical Bayesian space–time, mixed models, P–splines and conditional autoregressive models -CAR-, among others) of infectious disease counts, collected in different geographic areas, using multivariate or longitudinal data.

Functional data analysis (FDA) has grown in popularity over recent years alongside the increasing availability of continuous measurements in different contexts like Biomedicine [12], Spectrometry [13], Biology [14] and Medicine [15], to mention only a few. This study extends the regression models for independent functional data to the case where the curves presents either spatial or temporal dependencies.

Our goal is to estimate the rate of influenza epidemics, using the information readily available from public sources possibly that include functional variables, by adapting or extending the GLS techniques from a multivariate framework to this new framework. So, our particular aim is to estimate dependence components of influenza, using regression models, and predict the rate of incidence of influenza for a horizon of two weeks. We initially model influenza using a traditional linear approach (with independent errors) and later extend these ideas to the functional case (with dependent errors).

The article is structured as follows. Methodology section presents the Generalized Least Squares (GLS) approach for functional regression models. The estimation of the different parameters (for the regression function or the dependence) is usually done using maximum

likelihood although, as an alternative, we introduce an iterative GLS (iGLS) procedure that provides similar results. The latter could be interesting when the structure of the dependence is complicated. The practical performances of the GLS and iGLS procedures are compared, by means of a simulation study. Real example section applies these models to the prediction of the influenza rate in a region of Spain.

## Methodology

The functional regression model (FRM) is one of the most studied topics in FDA over the last few years. A regression model is said to be “functional” if any of the variates involved (the predictors or the response) has a functional nature, i.e. it is a measure observed along a continuous interval. Cases with a scalar response and functional predictors have particularly attracted a lot of attention. For example, Sørensen et al. [15] gives a basic introduction for the analysis of functional data applied in datasets from medical science.

The functional regression model with scalar response (FRM) is stated as follows: Let  $\mathcal{X}, y$  two random variates taking values in  $\mathcal{E} \times \mathbb{R}$  where  $\mathcal{E}$  is a functional space (semi-metric, normed or Hilbert). The relationship between the two variates can be expressed as follows:

$$y = m(\mathcal{X}) + \epsilon = \mathbb{E}(y|\mathcal{X}) + \epsilon \quad (1)$$

where  $\epsilon$  is a real random variable verifying  $\mathbb{E}(\epsilon|\mathcal{X}) = 0$ . Depending on the nature of the functional space  $\mathcal{E}$  and on the regression operator  $m$ , we can classify the different types of FRM:

- **Multivariate Linear Model:**  $\mathcal{E} = \mathbb{R}^p$  and  $m$  is the linear operator in the space, i.e.  $\mathbb{E}(y|\mathcal{X}) = \mathcal{X}\beta$  with  $\beta \in \mathbb{R}^p$ .
- **Functional Linear Model:**  $\mathcal{E} = \mathcal{L}^2(T)$  is the Hilbert space of square integrable functions over  $T = [a, b]$  and  $m$  is a linear operator in the space, i.e.  $m(\mathcal{X}) = \langle \mathcal{X}, \beta \rangle$  with  $\beta \in \mathcal{L}^2(T)$ . This model has been treated extensively in the literature mainly devoted to the optimal way of representing the linear operator through the representation of  $\mathcal{X}$  and  $\beta$  on a basis of  $\mathcal{L}^2(T)$ .

Depending on the latter, the references can be classified into two main categories:

- Fixed basis. The most commonly used basis in this context are the Fourier [16], the B-spline [17] and the Wavelet [18].
- Data-driven basis. Two main basis computed from the data are used in the literature: the most parsimonious one is given by the functional principal components [19, 20] and the one that maximizes the covariance among the response and the functional predictor uses the functional partial least square components (PLS) [21, 22].

Note that, due to the representation employed, the FRM is always an approximated model and its goodness typically relies on the properties of the chosen basis and its suitability to the data at hand.

- **Functional Non Linear Model:**  $\mathcal{E}$  is (at least) a semi-metric space and  $m$  is a continuous operator i.e.  $\lim_{\mathcal{X}' \rightarrow \mathcal{X}} m(\mathcal{X}') = m(\mathcal{X})$ . For a complete review of this model see Ferraty and Vieu [13] and the references therein.
- **Extensions of the above models:** The above models could be extended in several ways, usually considering more than one predictive variate. This could lead to semi-linear models [23, 24], additive models [25, 26], [27], single index models [28, 29] or projection pursuit models [30].

Many of the above-mentioned authors consider that  $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$  is an homoskedastic independent error vector, i.e.  $\mathbb{E}(\epsilon) = 0$ ,  $\text{Var}(\epsilon) = \sigma^2$  and  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ ,  $i \neq j$ . This assumption is made to obtain simple diagnostics or confidence intervals for the response but it could be too restrictive in functional regression models and difficult to check or fulfill in practice. Some papers consider dependence in the functional variate. See, for example, [31, 32] and [33] for contributions devoted to spatial dependence with functional data or [34–36] and [37] for time dependence. In both cases, the functional nature of the variate complicates the predictive ability of the model. The aim of this paper is to extend the GLS approach [38] to the functional context as the simplest way of incorporating temporal or spatial dependence in the regression models. In fact, the GLS approach can handle a wide range of regression models with dependence in a simple way: equi-correlation models, random effects, time and spatial dependence, and so on. This idea was first introduced in the context of FDA in [39].

## Functional generalized least squares regression

The functional generalized least squares regression (FGLS) model between two centered variables ( $\mathbb{E}(y) = 0$ ,  $\mathbb{E}(\mathcal{X}) = 0$ ) states that

$$y = \langle \mathcal{X}, \beta \rangle + \epsilon = \int_T \mathcal{X}(t)\beta(t)dt + \epsilon \quad (2)$$

where  $\beta \in \mathcal{L}_2(T)$  and  $\epsilon$  is now a random vector with mean 0 and covariance matrix  $\Omega = \mathbb{E}(\epsilon\epsilon')$ . This model includes, as its special cases, many others models, all of them based on  $\Omega = \Omega(\phi) = \sigma^2 \Sigma(\phi)$ , where  $\phi$  is the parameter associated with the dependence structure of  $\Omega$ . Some classical examples are presented in the following models:

1. Equi-correlated model:  $\text{Var}(\epsilon_i) = \sigma^2$  and  $\text{Cov}(\epsilon_i, \epsilon_j) = \sigma^2 \phi$ ,  $i \neq j$ ,  $\phi \in (-1, 1)$
2. Heteroskedastic block model:  $\Omega = \text{diag}(\sigma_1^2 \mathbf{I}_{n_1} | \sigma_2^2 \mathbf{I}_{n_2} | \dots | \sigma_p^2 \mathbf{I}_{n_p})$  with  $n_1 + n_2 + \dots + n_p = n$
3. AR(1) model:  $\epsilon_i = \phi \epsilon_{i-1} + \epsilon_i$  with  $|\phi| < 1$ ,  $\mathbb{E}(\epsilon_i) = 0$ ,  $\text{Var}(\epsilon_i) = \tau^2$  and  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ ,  $i \neq j$

$$\Omega = \frac{\tau^2}{1 - \phi^2} \left( \phi^{|i-j|} \right)_{i,j=1}^n$$

The variance structure is also known for every ARMA( $p, q$ ) model.

4. Spatial correlation model:

$$\Omega = \sigma^2 (\rho(d(s_i, s_j)))$$

where  $s_i, s_j$  are, respectively, the locations for  $i, j$ ; and  $\rho$  is the spatial correlation function.

**Estimation of functional GLS.** The classical theory of Kariya and Kurata [38] can be extended to the functional case by adapting the GLS criterion accordingly, i.e.

$$GLS(\beta, \phi) = (y - \langle \mathcal{X}, \beta \rangle)' \Sigma(\phi)^{-1} (y - \langle \mathcal{X}, \beta \rangle)$$

Given the sample  $\{(\mathcal{X}_1, y_1), \dots, (\mathcal{X}_n, y_n)\}$ , we can approximate  $\mathcal{X}_i$  and  $\beta$  using a finite sum of the basis elements:

$$\mathcal{X}_i(t) \approx \sum_k^{K_x} c_{ik} \psi_k(t), \quad \beta(t) \approx \sum_k^{K_\beta} b_k \varphi_k(t)$$

The preceding equations can be expressed as matrix notation using the evaluation in a grid of the length  $M$   $\{a = t_1 < \dots < t_M = b\}$  as

$$\mathbf{X} = \mathbf{C}\Psi, \quad \mathbf{B} = \mathbf{b}'\phi,$$

where  $\mathbf{X}$  is the matrix  $n \times M$  with the evaluations of the curves in the grid,  $\mathbf{C}$  is the matrix  $n \times K_x$  with the coefficients of the representation in the basis and  $\Psi$  is the matrix  $K_x \times M$  with the evaluations of the basis elements on the grid. Similarly,  $\mathbf{B}$  is the matrix  $(1 \times M)$  with the evaluation of the  $\beta$  parameter on the grid,  $\phi$  is the matrix  $(K_\beta \times M)$  with the evaluations of the basis  $\{\varphi_j\}$  and  $\mathbf{b}$  on the grid, is the vector of the coefficients of  $\beta$  in the basis.

With this notation, the terms  $\{\langle \mathcal{X}_i, \beta \rangle\}_{i=1}^n$  can be approximated by  $\mathbf{C}\Psi\phi'\mathbf{b} = \mathbf{Z}\mathbf{b}$  which, in essence, is a reformulation of a classical multivariate linear model that approximates the functional model. Here, the matrix  $\mathbf{Z}$  takes into account all the approximation steps done with the available information: the chosen basis for  $\mathcal{X}$  and  $\beta$  with the selected components:  $K_x$  and  $K_\beta$ .

Once a certain approximation is selected, supposing that  $\phi$  is known, we can define  $\mathbf{W} = \Sigma(\phi)^{-1}$ , and use the classical theory for multivariate GLS to obtain the BLUE of  $\mathbf{b}$  through:

$$\mathbf{b}_\Sigma = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}\mathbf{y},$$

where  $\mathbf{b}_\Sigma$  has covariance

$$\text{Cov}(\mathbf{b}_\Sigma) = \sigma^2 (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}$$

Finally, the fitted values are obtained by:

$$\hat{\mathbf{y}} = \mathbf{Z}(\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}\mathbf{y} = \mathbf{H}\mathbf{y}$$

where  $\mathbf{H}$  is the hat matrix.

Once the model is estimated, we can compute the prediction for a collection of  $m$  new data  $\{\mathcal{X}'_0\}$  using the model chosen for  $\Sigma$ . Being  $\epsilon_0$  the vector of errors for the new points,  $\Delta' = \text{Cov}(\epsilon, \epsilon_0)$  and  $\Sigma_0 = \mathbb{V}\text{ar}(\epsilon_0)$ , we can obtain the equations for prediction:

$$\hat{y}_0 = \langle \mathcal{X}_0, \hat{\beta} \rangle + \Delta \Sigma^{-1} (y - \langle \mathcal{X}, \hat{\beta} \rangle)$$

$$\mathbb{V}\text{ar}(\hat{y}_0) = \sigma^2 (\Sigma_0 - \Delta \Sigma^{-1} \Delta')$$

The GLS criterion can be employed to jointly estimate all the parameters associated to the model and can be expressed as:

$$\min_{K_x, K_\beta, \mathbf{b}, \phi} GLS = \min_{K_x, K_\beta, \mathbf{b}, \phi} (y - \mathbf{Z}\mathbf{b})' \Sigma(\phi)^{-1} (y - \mathbf{Z}\mathbf{b}),$$

where the parameters  $K_x$  and  $K_\beta$  related to the basis for  $\mathcal{X}$  and  $\beta$  are typically chosen *a priori* taking into account, for instance, the quality of the data and its representation on the discretization grid or other considerations related to the data-generating process (smoothness, physical restrictions, interpretability, . . .). The direct minimization of GLS usually cannot be affordable even though we only consider the parameters  $\mathbf{b}$  and  $\phi$ . The generalized cross-validation (GCV) criterion has been widely used to this end despite not being the right criterion for dependent errors. We use the generalized correlated cross-validation (GCCV) as a better alternative. This suggested criterion is an extension to GCV within the context of correlated errors

proposed by Carmack et al. [40]. It is defined as follows:

$$GCCV(K_x, K_\beta, \mathbf{b}, \phi) = \frac{\sum_{i=1}^n (y_i - \hat{y}_{i,b})^2}{\left(1 - \frac{\text{tr}(\mathbf{G})}{n}\right)^2}$$

where  $\mathbf{G} = 2\mathbf{H}\Sigma(\phi) - \mathbf{H}\Sigma(\phi)\mathbf{H}'$  takes into account the effect of the dependence, the trace of  $\mathbf{G}$  is an estimation of the degrees of freedom consumed by the model and  $\mathbf{H}$  is the hat matrix. The important advantage of this criterion is that it is rather easy to compute because it avoids the need to compute the inverse of the matrix  $\Sigma$ . Even so, the complexity of the GLS criterion depends on the structure of  $\Sigma$  and it could sometimes be hard either to minimize or computationally expensive.

We implement the function `fregre.gls` (and `predict.fregre.gls`) that estimates (and predicts) the functional regression model with correlated errors, see [S1 Appendix](#). The `fregre.gls` function calls the `gl`s function of `nlme` package. Therefore, the correlation structures allowed are those programmed by the original authors of the package [41].

**Estimation of functional iterative GLS (iGLS).** The above GLS criterion is employed to jointly estimate all the parameters associated with the model:  $K_x$ ,  $K_\beta$ ,  $\mathbf{b}$  and  $\phi$ . One possibility to alleviate the computational burden is to separate the estimation of the dependence structure ( $\phi$ ) from the parameters associated to the regression ( $K_x$ ,  $K_\beta$ ,  $\mathbf{b}$ ) in an iterative way (called iGLS) as it is done in multivariate regression. The iGLS proven to be equivalent to classical GLS (see, for instance, [42]). Additionally, the method could consider more flexible dependence models (for instance, selecting the order of an AR instead of fixing it in advance) that avoid the risk of misspecification in the dependence structure. We extend this procedure to the functional regression in the following iterative procedure (called functional iGLS):

1. Begin with a preliminary estimation of  $\hat{\phi} = \phi_0$  (for instance,  $\phi_0 = 0$ ). Compute  $\hat{\mathbf{W}}$ .
2. Estimate  $\mathbf{b}_\Sigma = (\mathbf{Z}'\hat{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{W}}\mathbf{y}$
3. Based on the residuals,  $\hat{e} = (\mathbf{y} - \mathbf{Z}\mathbf{b}_\Sigma)$ , update  $\hat{\phi} = \rho(\hat{e})$  (and consequently,  $\hat{\mathbf{W}}$ ) where  $\rho$  is subject to the dependence structure chosen.
4. Repeat steps 2 and 3 until convergence (small changes in  $\mathbf{b}_\Sigma$  and/or  $\hat{\phi}$ )

The estimation of functional  $\beta(t)$  by  $\mathbf{b}_\Sigma$  is done in step (2), and separated from the estimation of dependence structure  $\rho$  in step (3). This allows for the flexibility of including any type of dependence structures designed by the user (for instance, using particular restrictions) that are typically not included in the usual packages (like `nlme`).

We implement, the function `fregre.igls` (and `predict.fregre.igls`) that estimates (and predicts) the functional regression model with correlated errors using the iterative scheme (iGLS). We have developed the following two simple structures for  $\Sigma$  in `fda.usc` package [45] for fit serial dependence structure:

- In iGLS-AR( $p$ ) scheme, the procedure automatically fits the autoregressive order  $p$  in each iteration of the errors defined by the equation  $\epsilon_i = \sum_{j=1}^p \phi_j \epsilon_{i-j} + \epsilon_i$  where  $\epsilon_i \sim N(0, \sigma^2)$ .
- In iGLS-ARMA( $p, q$ ) scheme, the user must specify the parameters  $p$  and  $q$  of the autoregressive-moving-average (ARMA( $p, q$ )) model, which fits the serial error dependence defined by equation:  $\epsilon_i = \sum_{j=1}^p \phi_j \epsilon_{i-j} + \sum_{j=1}^q \theta_j \epsilon_{i-j} + \epsilon_i$  where  $\epsilon_i \sim N(0, \sigma^2)$ . This structure is provided by the `nlme` package but it has a restriction: all parameters of the AR side must be lower

than one in absolute value. This rule clearly does not include all the possible stationary models of that order (this is only true for ARMA(1,q)).

For these structures, we have used the basic functions `ar` and `arima` of the `stats` package to fit the AR( $p$ ) and ARMA( $p,q$ ) models, respectively. The users can define their own functions or use other well-known functions that exactly fit the situation at hand.

## Simulation

We have used two functional linear models (FLM) included in [17] to compare the effect of the temporal dependence. Specifically, we have generated  $n_B = 1000$  replicas of size  $n = 100$  from the FLM model  $y = \langle \mathcal{X}, \beta \rangle + \epsilon$ , being  $\mathcal{X}$  a Wiener process observed in a grid of length  $M = 100$  in the interval  $[0, 1]$  and  $\epsilon$  an AR(1) process with autoregressive parameter  $\phi$  and variance  $\text{Var}(\epsilon) = \text{snr} \text{Var}(\langle \mathcal{X}, \beta \rangle)$ , where  $\text{snr}$  is the signal to noise ratio. For each sample, ten future values, denoted by  $(y_{n+h}, h = 1, \dots, 10)$ , were generated to check the predictive ability of the proposal.

The two models differ only in the  $\beta$  parameter that are respectively:

1.  $\beta(t) = 2 \sin(0.5\pi t) + 4 \sin(1.5\pi t) + 5 \sin(2.5\pi t)$ ,  $t \in [0, 1]$ ,
2.  $\beta(t) = \log(15t^2 + 10) + \cos(4\pi t)$ ,  $t \in [0, 1]$ .

The scenario (a) corresponds to a  $\beta$  parameter which has an exact representation respect to the first three theoretical principal components of the Wiener process. On the contrary, the  $\beta$  parameter for scenario (b) cannot be well represented using a small number of theoretical principal components. In both scenarios, we have used two types of basis for representing  $\mathcal{X}$  and  $\beta$ : the empirical principal components basis derived from the sample (FPC) and the cubic B-splines (BSP) at equispaced knots in  $[0, 1]$ . The same basis was employed for both representations i.e. in this case  $\Psi = \varphi$  and  $K_x = K_\beta$ . The optimal number of components ( $K_\beta$ ) was selected using the GCCV criterion in the range 1–8 for FPC and 5–11 for BSP.

For sake of simplicity, we only show here the results for model (a). The results for model (b) can be revised in the S2 Appendix of the Supporting information.

Tables 1 to 4 summarize the results for the first model (a) to show, respectively, the average number of selected components chosen using GCCV criterion, the mean square error (MSE) for estimation of  $\beta$ , the MSE for estimation of  $\phi$  and the mean square prediction errors (MSPE) for horizons 1, 5 and 10. In these results, LM denotes the estimation through a classical functional linear model whereas GLS and iGLS corresponds, respectively, to the functional GLS and functional iGLS methods (shown in Methodology section for AR(1) dependent errors).

$$\mathbb{E} \left( \left\| \beta - \hat{\beta} \right\|^2 \right)$$

$$\mathbb{E} \left( \left( \phi - \hat{\phi} \right)^2 \right)$$

**Table 1. Average of number of basis elements selected by GCCV criterion in  $n_B = 1000$  replicas in model (a).**

snr	PC			BSP		
	$\phi = 0$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0$	$\phi = 0.5$	$\phi = 0.9$
0.05	3.3	3.3	3.3	6.4	6.5	6.6
0.10	3.2	3.2	3.3	5.9	6.0	6.1

<https://doi.org/10.1371/journal.pone.0194250.t001>



**Table 2. Mean square error of  $\beta$  parameter.** Model (a),  $n_B = 1000$ .

snr	Model	PC			BSP		
		$\phi = 0$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0$	$\phi = 0.5$	$\phi = 0.9$
0.05	LM	0.457	0.442	0.443	0.996	1.014	0.965
0.05	GLS-AR(1)	0.457	0.421	0.400	0.997	0.813	0.493
0.05	iGLS-AR(1)	0.457	0.421	0.400	0.997	0.813	0.493
0.05	iGLS-AR(p)	0.457	0.421	0.400	1.001	0.816	0.494
0.10	LM	0.501	0.502	0.503	1.243	1.261	1.218
0.10	GLS-AR(1)	0.501	0.471	0.437	1.244	1.031	0.661
0.10	iGLS-AR(1)	0.501	0.471	0.437	1.244	1.031	0.661
0.10	iGLS-AR(1)	0.501	0.471	0.437	1.247	1.032	0.662

<https://doi.org/10.1371/journal.pone.0194250.t002>

**Table 3. Mean square error of  $\phi$  parameter.** Model (a),  $n_B = 1000$ .

snr	Model	PC			BSP		
		$\phi = 0$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0$	$\phi = 0.5$	$\phi = 0.9$
0.05	GLS-AR(1)	0.004	0.003	0.001	0.004	0.003	0.001
0.05	iGLS-AR(1)	0.004	0.003	0.002	0.004	0.003	0.001
0.10	GLS-AR(1)	0.004	0.003	0.001	0.004	0.003	0.001
0.10	iGLS-AR(1)	0.004	0.003	0.001	0.004	0.003	0.001

<https://doi.org/10.1371/journal.pone.0194250.t003>

$$MSPE = \frac{1}{n_B} \sum_{b=1}^{n_B} (y_{n+h}^b - \hat{y}_{n+h}^b)^2$$

Table 1 shows an average number of FPC selected components between 3 and 4 with a slight tendency to lower values as the *snr* grows. The average number of B-splines basis was

**Table 4. Mean square prediction errors for lags  $h = 1, 5$  and  $10$ .** Model (a),  $n_B = 1000$ .

snr	Model	Basis	AR(1)								
			$\phi = 0$			$\phi = 0.5$			$\phi = 0.9$		
			$h = 1$	$h = 5$	$h = 10$	$h = 1$	$h = 5$	$h = 10$	$h = 1$	$h = 5$	$h = 10$
0.05	LM	PC	0.070	0.068	0.065	0.070	0.077	0.068	0.070	0.070	0.070
0.05	GLS-AR(1)	PC	0.070	0.068	0.065	0.054	0.077	0.068	0.015	0.047	0.061
0.05	iGLS-AR(1)	PC	0.070	0.068	0.065	0.054	0.077	0.068	0.015	0.047	0.061
0.05	iGLS-AR(p)	PC	0.070	0.068	0.065	0.055	0.076	0.068	0.015	0.047	0.062
0.05	LM	BSP	0.071	0.069	0.066	0.071	0.077	0.068	0.071	0.072	0.070
0.05	GLS-AR(1)	BSP	0.071	0.069	0.066	0.054	0.076	0.068	0.015	0.046	0.060
0.05	iGLS-AR(1)	BSP	0.071	0.069	0.066	0.054	0.076	0.068	0.015	0.046	0.060
0.05	iGLS-AR(p)	BSP	0.072	0.069	0.066	0.055	0.076	0.068	0.015	0.047	0.061
0.10	LM	PC	0.137	0.152	0.126	0.150	0.137	0.141	0.136	0.136	0.153
0.10	GLS-AR(1)	PC	0.137	0.152	0.126	0.114	0.137	0.140	0.030	0.094	0.136
0.10	iGLS-AR(1)	PC	0.137	0.152	0.126	0.114	0.137	0.140	0.030	0.093	0.136
0.10	iGLS-AR(p)	PC	0.138	0.152	0.126	0.115	0.138	0.140	0.030	0.094	0.136
0.10	LM	BSP	0.138	0.153	0.130	0.150	0.137	0.143	0.140	0.137	0.155
0.10	GLS-AR(1)	BSP	0.139	0.153	0.130	0.114	0.137	0.140	0.029	0.094	0.136
0.10	iGLS-AR(1)	BSP	0.139	0.153	0.130	0.114	0.137	0.140	0.029	0.093	0.135
0.10	iGLS-AR(p)	BSP	0.140	0.153	0.130	0.115	0.137	0.140	0.030	0.094	0.136

<https://doi.org/10.1371/journal.pone.0194250.t004>



between 6 and 7 although in this case we do not have a theoretical quantity to compare with. It seems that there are no trends with respect to the  $\phi$  values. Table 2 clearly shows the advantage of the PC estimator over the B-splines because the estimation error using B-splines typically doubles the error using PCs. In this table, we can also see the improved estimates of the GLS and iGLS method over the LM, especially when  $\phi$  grows. The same equivalence is shown in Table 3 for the mean square error (MSE) of the  $\phi$  parameter, which shows better results as the dependence grows. Finally, Table 4 shows the mean square prediction errors (MSPE) for different lags showing a clear improvement of GLS procedures, specially for large  $\phi$  and shorter lags. With respect to the prediction ability between PC or B-splines, the results show that both methods are almost equivalent with minor differences along the table.

Table 5 summarizes the results of the Model (a) but replaces the AR(1) by an AR(2) error process using the FPC estimation (the results with BSP are similar). In all these models, the minimum square prediction error is achieved with model iGLS-AR(2) in which an AR(2) is estimated in each iteration of the algorithm. This is followed very closely by model iGLS-AR( $p$ ), estimating an automatic choice of  $p$  at each iteration.

$$MSPE = \frac{1}{n_B} \sum_{b=1}^{n_B} (y_{n+h}^b - \hat{y}_{n+h}^b)^2$$

The first AR(2) process, ( $\phi_1 = 0.5, \phi_2 = 0.45$ ), is roughly like an AR(1) process with  $\phi \approx 0.95$ . This can explain why the results of the iGLS-AR(1) model are so close to the optimum estimated by the iGLS-AR(2). The second AR(2) process, ( $\phi_1 = 1.4, \phi_2 = -0.45$ ), was selected to assess the misspecification error. Although the use of an AR(1) process in the GLS and iGLS models improves the LM model, these results are far from the best using an AR(2) specification. The autocorrelation function of the AR(2) process shows a periodicity pattern that cannot be approximated by an AR(1) process. Finally, the third AR(2), ( $\phi_1 = 1.5, \phi_2 = -0.75$ ), shows the effect of the misspecification in a later horizon  $h = 5$ , making the results at that horizon for an AR(1) specification even worse than the LM model. Again, this is motivated by the periodicity pattern of the AR(2) due to the negative sign of  $\phi_2$ . In all cases, the specification iGLS-AR( $p$ ) is rather close to the optimum. However, the important advantage is that it avoids a closed specification form of the dependence structure. Finally, the GLS-AR(2) scenario was not considered in this table because the `gls` function of `nlme` package does not allow the estimation of any parameter of an AR(2) greater than 1 in absolute value. This is an empirical rule in the package that avoids the use of non stationary processes although, in this case, the three AR(2) specifications are clearly stationary, but only the first specification can be estimated using the `gls` function.

**Table 5. Mean square prediction errors for different lags  $h = 1, 5, 10$ . Estimation of Model (a) using PC with an AR(2) error process.**

snr	Model	AR(p = 2)								
		$(\phi_1 = 0.5, \phi_2 = 0.45)$			$(\phi_1 = 1.4, \phi_2 = -0.45)$			$(\phi_1 = 1.5, \phi_2 = -0.75)$		
		$h = 1$	$h = 5$	$h = 10$	$h = 1$	$h = 5$	$h = 10$	$h = 1$	$h = 5$	$h = 10$
0.05	LM	0.0707	0.0693	0.0725	0.0643	0.0628	0.0667	0.0657	0.0688	0.0737
0.05	GLS-AR(1)	0.0144	0.0304	0.0490	0.0154	0.0493	0.0626	0.0191	0.1072	0.0714
0.05	iGLS-AR(1)	0.0144	0.0309	0.0497	0.0154	0.0483	0.0600	0.0191	0.1055	0.0712
0.05	iGLS-AR(2)	0.0109	0.0250	0.0415	0.0050	0.0358	0.0551	0.0092	0.0572	0.0682
0.05	iGLS-AR(p)	0.0115	0.0256	0.0424	0.0052	0.0364	0.0554	0.0093	0.0569	0.0682

<https://doi.org/10.1371/journal.pone.0194250.t005>

## An application to Galician flu prediction rate

Galicia is a region of 29,574 km<sup>2</sup> located in Northwest Spain with a population of 2.8 million people. We analyzed the weekly incidence of reported cases of influenza in Galicia between 2001 and 2011 for each of the 53 Galician counties:

$$\text{Rate}_{n,s} = \log(\text{cases}_{n,s} \times 100000 / \text{pop}_{n,s})$$

for county  $s$  and week  $n$ . The population ( $\text{pop}$ ) was obtained from the Statistical Institute of Galicia (IGE, <http://www.ige.eu>) and the number of influenza cases ( $\text{cases}$ ) from the Health Service of Galicia ([www.sergas.es](http://www.sergas.es)).

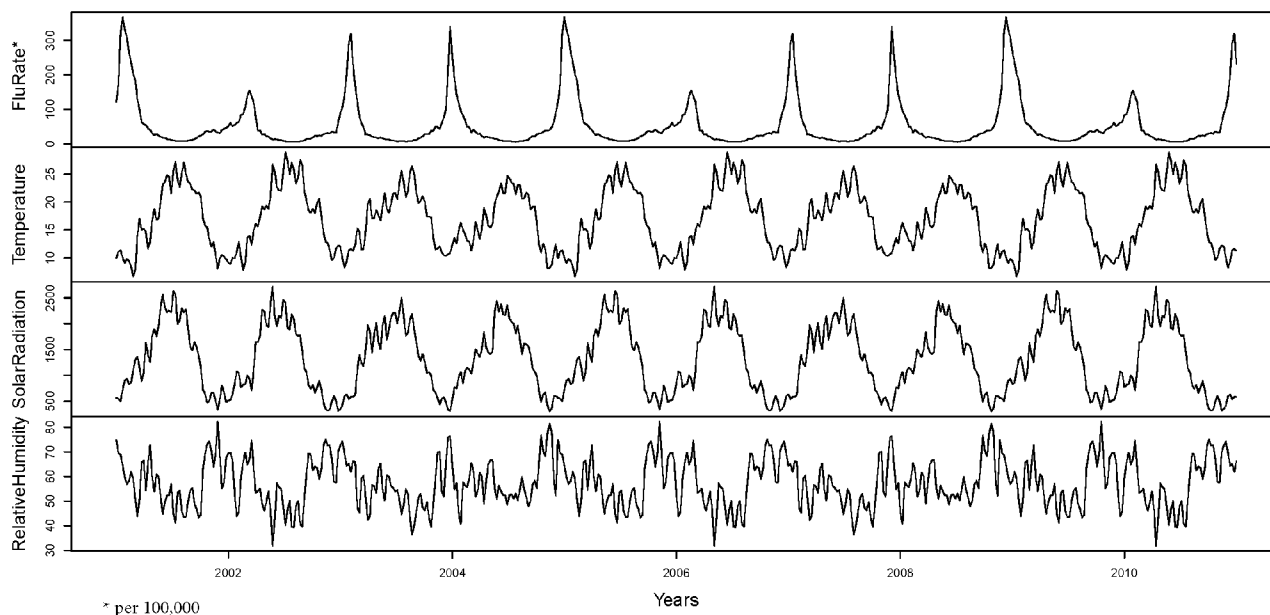
The influenza season in Galicia usually begins in week 40 and ends in week 20 of the following year. The goal is to predict the incidence of influenza for the following two weeks ( $n+1$  and  $n+2$ ) for each of the  $s$  regions with the available information:

- $\text{Rate}_{n,s}(w)$ : Weekly influenza rate for last 13 weeks,  $w \in [n-12, n]$ .
- $\text{Temp}_{n,s}(t)$ : Daily temperature in Celsius degrees ( $^{\circ}\text{C}$ ) for last 14 days,  $t \in [n-i/7, n]$ , for  $i = 14, \dots, 1$ .
- Dushoff et al. [7] defined cold as the number of degrees below a threshold temperature:  $\text{Temp.th}_{n,s} = \min(\text{Temp}_{n,s} - \text{thres}, 0)$  with  $\text{thres} = 10^{\circ}\text{C}$ . The functional variable is defined as:  $\text{Temp.th}_{n,s}(t)$  with  $t \in [n-i/7, n]$ , for  $i = 14, \dots, 1$ .
- $\text{SR}_{n,s}(t)$ : Daily solar radiation ( $\text{W/m}^2$ ) for the last 14 days,  $t \in [n-i/7, n]$ , for  $i = 14, \dots, 1$ .
- $\text{Hum}_{n,s}(t)$ : Relative humidity for the last 14 days:  $t \in [n-i/7, n]$ , for  $i = 14, \dots, 1$ .

For representing the above functional covariates, a B-spline basis of five components was used in all cases (based on the previous experience of the authors with this type of data). The prediction for the overall influenza rate is constructed by appropriately aggregating the predictions of the  $s$  regions that are made independently, i.e. the estimation of  $\beta$  and  $\phi$  are made only with the data of that county. Fig 1 shows the overall influenza rate that normally grows in the late autumn and reaches a peak at the beginning of the calendar year. These plots clearly show the large difference between reported influenza cases in winter and summer. The influenza rate for each county shows a similar pattern but with small differences in the peak epidemic period. We downloaded meteorological data from the regional Weather Service of Galicia (<http://www.meteogalicia.es/>). S1 Appendix describes the supplementary material (functions, libraries, source data and code) and S1 File contains the code and dataset used in this study.

## Variable selection using distance correlation measure

Distance correlation  $\mathcal{R}$  is a measure of dependence between random vectors introduced by Székely et al. [43]. The distance correlation satisfies  $0 \leq \mathcal{R}(X, Y) \leq 1$  and its interpretation is similar to the squared Pearson's correlation. However, the advantages of distance correlation over the Pearson correlation is that it defines  $\mathcal{R}(X, Y)$  in arbitrary finite dimensions of  $X$  and  $Y$  and  $\mathcal{R}$  characterises independence, i.e.  $\mathcal{R}(X, Y) = 0 \Leftrightarrow X, Y$  are independent. Recently, Lyons [44] provided conditions for the application of the distance correlation to functional spaces. So, this measure seems to be a good indicator of the correlations between functional and multivariate variables that may be useful for designing a functional linear model (for instance, avoiding variates with high collinearity). The empirical distance correlation



**Fig 1. Influenza rate and meteorological covariates.** From top to bottom: Overall weekly influenza rate, and daily average temperature, solar radiation and relative humidity in the Galician region during the period.

<https://doi.org/10.1371/journal.pone.0194250.g001>

$\mathcal{R}_{n,s}(\mathbf{X}, \mathbf{Y})$  can be easily computed as

$$\mathcal{R}_{n,s}(\mathbf{X}, \mathbf{Y}) = \sqrt{\frac{\mathcal{V}_{n,s}^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{V}_{n,s}^2(\mathbf{X})\mathcal{V}_{n,s}^2(\mathbf{Y})}}}$$

where  $\mathcal{V}_{n,s}(\mathbf{X}, \mathbf{Y})$  is the empirical distance covariance defined by

$$\mathcal{V}_{n,s}^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^n A_{kl} B_{kl}$$

where  $A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..}$  and  $B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..}$  with  $a_{kl} = \|X_k - X_l\|$ ,  $b_{kl} = \|Y_k - Y_l\|$ ,  $k, l = 1, \dots, n$ , and the subscript. denotes that the mean is computed for the index that it replaces. Similarly,  $\mathcal{V}_{n,s}(\mathbf{X})$  is the non-negative number defined by  $\mathcal{V}_{n,s}^2(\mathbf{X}) = \mathcal{V}_{n,s}^2(\mathbf{X}, \mathbf{X}) = \frac{1}{n^2} \sum_{k,l=1}^n A_{kl}^2$ .

The distance correlation  $\mathcal{R}$  was used to select the information relevant to the prediction of influenza rate not only with respect to the response but also among the possible covariates to avoid collinearities. The results are shown in Table 6. Relative humidity,  $\text{Hum}_{n,s}(t)$ , has the lowest correlation with the influenza rate  $\{\text{Rate}_{n+1,s}, \text{Rate}_{n+2,s}\}$  and therefore, it seems that its

**Table 6. Distance correlation  $\mathcal{R}$  between the response at week  $n+1$  and  $n+2$  and functional covariates at week  $n$ .**

$\mathcal{R}$	$\text{Rate}_{n,s}(w)$	$\text{Temp}_{n,s}(t)$	$\text{Temp.th}_{n,s}(t)$	$\text{SR}_{n,s}(t)$	$\text{Hum}_{n,s}(t)$	$\text{Rate}_{n+1,s}$	$\text{Rate}_{n+2,s}$
$\text{Rate}_{n,s}(w)$	1.00	0.56	0.48	0.43	0.26	0.69	0.64
$\text{Temp}_{n,s}(t)$	0.56	1.00	0.90	0.78	0.54	0.52	0.50
$\text{Temp.th}_{n,s}(t)$	0.48	0.90	1.00	0.73	0.44	0.46	0.45
$\text{SR}_{n,s}(t)$	0.43	0.78	0.73	1.00	0.72	0.52	0.51
$\text{Hum}_{n,s}(t)$	0.26	0.54	0.44	0.72	1.00	0.31	0.30

<https://doi.org/10.1371/journal.pone.0194250.t006>

contribution to the response is negligible (a model with  $\text{Hum}_{n,s}(t)$  never improves one without the variate). Besides, the distance correlation values are useful for designing models avoiding closely related covariates (for instance,  $\text{Temp}_{n,s}(t)$  and  $\text{Temp.th}_{n,s}(t)$  share the same information). With these considerations, the number of possible different models to be tested is quite reduced.

## Prediction using temporal dependence structure

A rolling analysis was employed to compare the models in a predictive scenario. Initially, a series of length  $j = 1, \dots, n = 150$  weeks in  $s = 53$  counties is used to predict the influenza rate in the next two weeks,  $n + 1$  and  $n + 2$ . The rolling is then performed along the epidemic periods ( $J = 28$  weeks, from week 40 to week 15 next year) by computing the mean square predictive error:

$$MSPE = \frac{1}{J} \sum_{j=n+1}^{n+J} \sum_{r=1}^s w_r \left( \text{Rate}_{j,r} - \widehat{\text{Rate}}_{j,r} \right)^2$$

where  $w_r$  is the weight (in terms of *pop*) for county  $r$ . For ease of simplicity, the GLS setting is only considered with an AR(1) specification of the dependence structure, whereas the iGLS is combined with an AR(1), AR(2) and AR( $p$ ).

Table 7 summarises the MSPE for the influenza season. The best result for each set of covariates is shaded in light gray and the overall winner for each horizon is in bold font. In the models with the predictor  $\text{Rate}_{n,s}(w)$  (rows (a), (e), (f) and (g)) the gain, in terms of MSPE, of the functional GLS models (GLS-AR(1), iGLS-AR(1), iGLS-AR(2) and iGLS-AR( $p$ )) is relatively small with respect to functional LM models because the  $\text{Rate}_{n,s}(w)$  partly accounts for the temporal dependence. Furthermore, in some sense, the inclusion of the predictor  $\text{Rate}_{n,s}(w)$  in the model is akin to the estimation of the dependence structure. The models without influenza

Table 7. Mean square predictive error for influenza period using the rolling procedure.

	$n + 1$				
Covariates	LM	GLS-AR(1)	iGLS-AR(1)	iGLS-AR(2)	iGLS-AR( $p$ )
(a) $\text{Rate}_{n,s}(w)$	0.510	0.404	0.404	0.405	0.402
(b) $\text{Temp}_{n,s}(t)$	1.177	<b>0.362</b>	<b>0.362</b>	0.364	0.379
(c) $\text{Temp.th}_{n,s}(t)$	2.530	0.391	0.391	0.402	0.418
(d) $\text{SR}_{n,s}(t)$	1.290	0.381	0.381	0.394	0.407
(e) $\text{Rate}_{n,s}(w), \text{Temp}_{n,s}(t)$	0.487	0.404	0.404	0.392	0.390
(f) $\text{Rate}_{n,s}(w), \text{Temp.th}_{n,s}(t)$	0.538	0.362	0.448	0.441	0.437
(g) $\text{Rate}_{n,s}(w), \text{SR}_{n,s}(t)$	0.505	0.418	0.418	0.409	0.404
(h) $\text{Temp}_{n,s}(t), \text{SR}_{n,s}(t)$	1.163	0.384	0.384	0.389	0.402
	$n + 2$				
Covariates	LM	GLS-AR(1)	iGLS-AR(1)	iGLS-AR(2)	iGLS-AR( $p$ )
(a) $\text{Rate}_{n,s}(w)$	0.931	0.903	0.901	0.849	0.809
(b) $\text{Temp}_{n,s}(t)$	1.250	0.785	0.764	0.760	<b>0.712</b>
(c) $\text{Temp.th}_{n,s}(t)$	1.954	0.841	0.830	0.828	0.792
(d) $\text{SR}_{n,s}(t)$	1.272	0.821	0.823	0.834	0.795
(e) $\text{Rate}_{n,s}(w), \text{Temp}_{n,s}(t)$	0.883	0.879	0.814	0.810	0.764
(f) $\text{Rate}_{n,s}(w), \text{Temp.th}_{n,s}(t)$	0.911	0.785	0.874	0.856	0.800
(g) $\text{Rate}_{n,s}(w), \text{SR}_{n,s}(t)$	0.951	0.939	0.880	0.877	0.855
(h) $\text{Temp}_{n,s}(t), \text{SR}_{n,s}(t)$	1.273	0.796	0.784	0.783	0.746

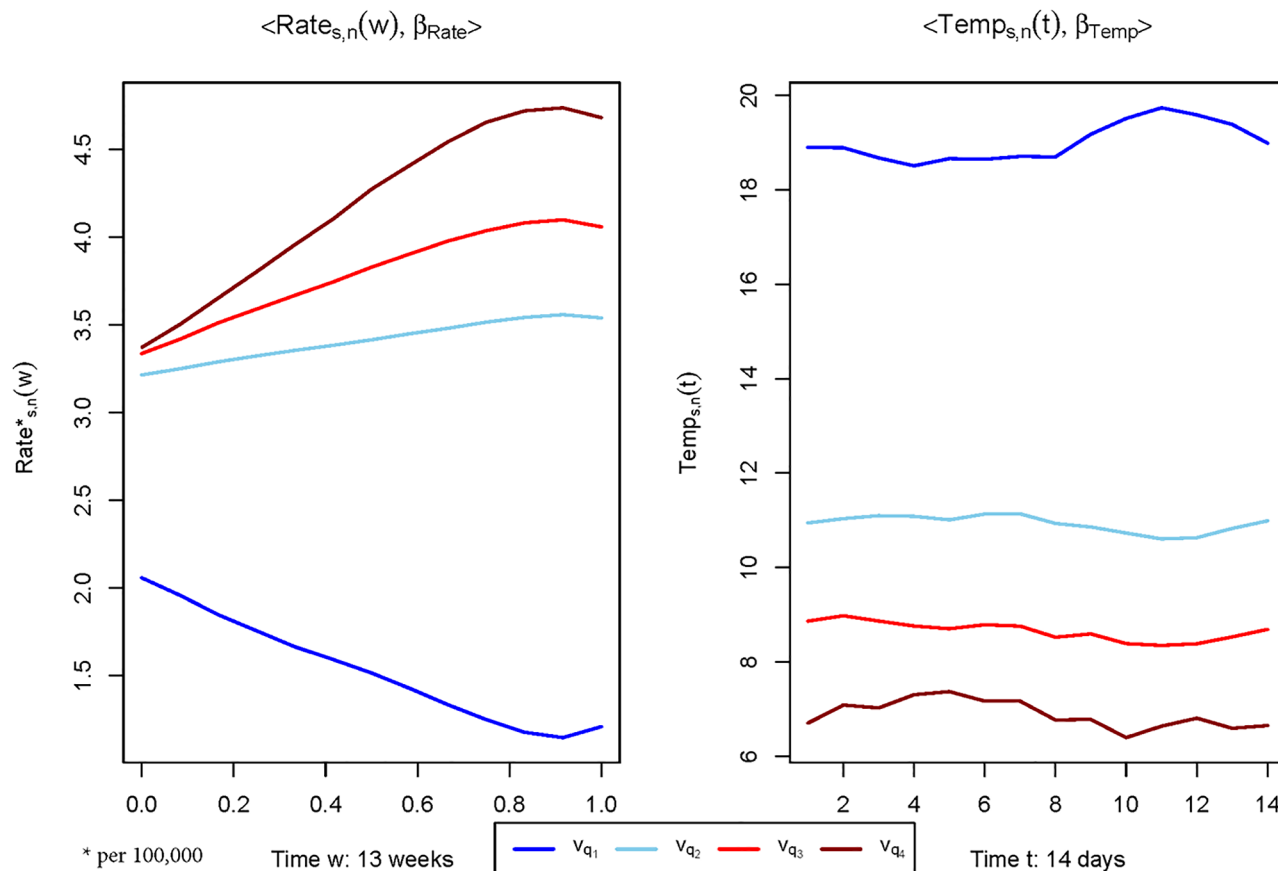
<https://doi.org/10.1371/journal.pone.0194250.t007>

rate (rows (b), (c), (d) and (h)) begin with a worse result in the LM setting, but their results become competitive (or even become the best ones) with the inclusion of the serial dependence. The difference between the GLS or iGLS setting is that the latter allows more flexibility, not only defining a different dependence structure in each county, but also in the estimation of that dependence. This is particularly useful when the forecast horizon increases. The GLS setting must fix the order of the AR in advance and, when the number of regions is high, it is a tough assumption to consider the order of the serial dependence model fixed for all of them. For  $n + 1$  the best models are (b) and (c) with GLS-AR(1) and iGLS-AR(1) specifications, using the curve of temperature of last 14 days as the predictor and a simple AR(1) structure for the adjustment of the residuals. The best autoregressive model estimated by the iGLS-AR( $p$ ) model has been, in most cases, of order 1. For  $n + 2$ , in some regions, an AR(1) or AR(2) model may be insufficient; the best result is achieved with the iGLS-AR( $p$ ) procedure, which presents greater flexibility in estimating the different  $p$  order for each county.

$$MSPE = \frac{1}{J} \sum_{j=n+1}^{n+J} \sum_{r=1}^s w_r \left( Rate_{j,r} - \widehat{Rate}_{j,r} \right)^2$$

Models (b) and (c), with GLS setting, present slight differences. Of course, it seems better to use the temperature than to only use the threshold respect to a level. Yet the differences between these two models suggest that the evolution of temperatures when it is cold is crucial to explaining the influenza rate. Model (h) makes no improvement on the results of models (b) and (c) in terms of MSPE. In fact, it worsens them; this is probably due to collinearity among  $Temp_{n,s}(t)$  and  $SR_{n,s}(t)$ . Concerning models (b), (c) and (d), the first two are preferable because they are easier to apply and interpret. Besides, in model (d) the measures of solar radiation usually depend on specialised devices, whereas the covariates related to temperature are readily available using standard (and cheaper) equipment. Finally, for short horizons, it seems unnecessary to specify high order autoregressive models, even though the improvement can be about 5% for larger lags.

Indeed, it is possible to interpret the  $\hat{\beta}$  parameter associated with models. To this end, we have computed for models (a) and (b), the quantities  $v_i = \langle \mathcal{X}_i, \hat{\beta} \rangle$ , which are the contribution of every curve to the influenza rate. So, if we classify the curves in groups according to these values and average them, we can see the pattern of the curves that have the most (or least) influence with respect to the incidence rate. This is done in Fig 2, which shows the pattern of curves that most contributed to increasing (in red scale) and decreasing (in blue scale) the influenza rate. In particular, we have split the data with respect to the quartiles of  $v_i$  and assigned (from bottom to top) the following colors: blue, sky blue, red and dark red. This assesses the evaluation of the contribution of these curves in the response. So, as expected, the contribution of an intense increasing pattern of the influenza rate in the last weeks is plotted in dark red (see left panel of Fig 2), which leads to predicting high influenza rates. On the other hand, a decreasing pattern is plotted in dark blue, meaning that this type of pattern corresponds with low influenza rates. The same reasoning can be applied to model (b) (see right panel of Fig 2). Curves of temperature below 7°C are plotted in dark red, meaning that this pattern provides high prediction rates. On the other hand, the curves around 19°C (plotted in dark blue) lead to almost zero influenza rates. The dark red line corresponds to the pattern of the curves that most contribute to increasing the estimated incidence rate. In the week  $w = 1$  begins  $v_{q_4} \approx 3.3$  that, if we undo the logarithmic transformation represents an incipient incidence of 27.1 cases per 100, 000 population and goes up monotonously until

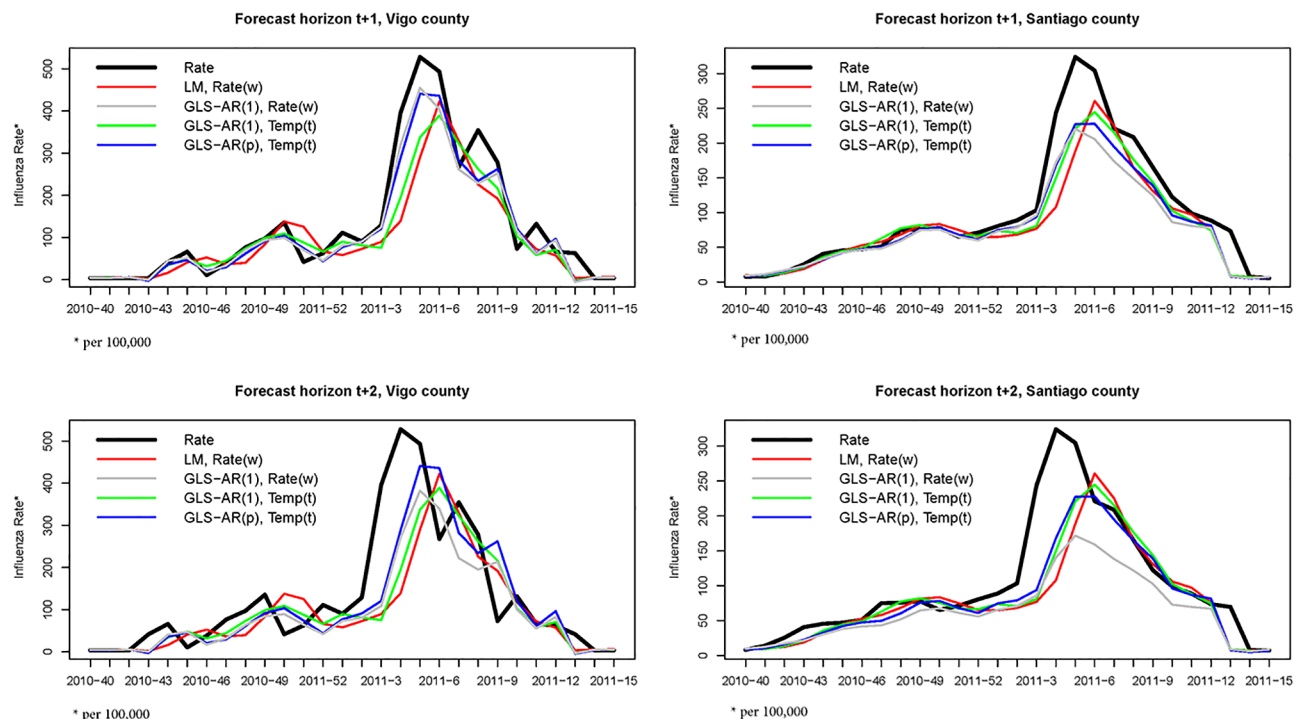


**Fig 2. Shape of the covariates with respect to their contribution in the model.** Shape of rate curves (on left) and temperature threshold curves (on right) categorized by their projection value  $v_X = \langle X, \hat{\beta} \rangle$ . The groups are constructed as a function of the quantile of  $v_X$  ( $q(v_X)$ ):  $q(v_X) \in [0, .25]$  (dark blue line),  $q(v_X) \in (0.25, 0.50]$  (blue line),  $q(v_X) \in (0.50, 0.75]$  (red line) and  $q(v_X) \in (0.75, 1]$  (dark red line).

<https://doi.org/10.1371/journal.pone.0194250.g002>

last register ( $w = 13$ ), which takes the value  $v_{q_1} \approx 4.6$ , implying an increase of 99.5 cases per 100,000 population.

Finally, as an illustration in Fig 3 the prediction of the raw rate ( $\text{cases} \times 100000/\text{pop}$ ) during the 2010–11 flu epidemic season is provided for two counties (Vigo and Santiago) as a result of reversing the log transform of the response in the preceding models. In both counties, the peak is achieved at week 2011–5 (first week of February). The two considered horizons ( $t + 1$  and  $t + 2$ ) are shown by rows. In each case, the raw rate is compared with the prediction obtained one or two weeks before with the models LM,  $\text{Rate}(w)$ ; GLS–AR(1),  $\text{Rate}(w)$ ; GLS–AR(1),  $\text{Temp}(t)$  and GLS–AR( $p$ ),  $\text{Temp}(t)$ . Focusing on  $t + 1$ , the comparison among the two dependence structures (AR(1) and AR( $p$ ), lines green and blue, respectively) associated with  $\text{Temp}(t)$  shows a big difference for Vigo but no for Santiago. This suggests that for Santiago an AR(1) is enough whereas for Vigo it seems more adequate a general AR( $p$ ) specification. Respect to the models including the  $\text{Rate}(t)$  (lines red and gray), the model using GLS reacts faster than the LM model providing better predictions of the peak. Predictions for medium or low intensities (below 125) are quite similar. For  $t + 2$ , no clear patterns are shown, although the specification GLS–AR( $p$ ) seems to do slightly better.



**Fig 3. Example of raw influenza rate prediction.** Prediction of the raw rate ( $\text{cases} \times 100000/\text{pop}$ ) for two counties (Vigo and Santiago) in Galicia using four models: LM,  $\text{Rate}(w)$ ; GLS-AR(1),  $\text{Rate}(w)$ ; GLS-AR(1),  $\text{Temp}(t)$  and GLS-AR(p),  $\text{Temp}(t)$ . In each case, the raw rate is compared with the prediction provided one week before ( $t+1$ , first row) and two weeks before ( $t+2$ , second row). The counties are separated by columns.

<https://doi.org/10.1371/journal.pone.0194250.g003>

## Conclusion

This paper extends the GLS model from a multivariate to a functional framework: it thereby allows us to estimate functional regression models with temporal or spatial covariance errors structure in a simple way. It proposes an iterative version of the GLS estimator, that can help to model very complicated dependence structures. This procedure (called iGLS) is much simpler than GLS in terms of the optimization function to be accomplished but, of course, it may take longer due to the iterations. However, iGLS may be the only option when the sample size or the dimension of the parameter increases and the joint optimization performed by GLS is not affordable (in terms of complexity or memory consumption).

A simulation study shows that the GLS estimators improve the classical approach because they provide better estimations of the parameters associated with the regression model and extremely good results from the predictive point of view, specially for short lags.

The GLS procedures have been applied to the prediction of the influenza rate using readily available functional variables. These kinds of models are extremely useful to health managers in allocating resources in advance for an epidemic outbreak. The estimation of the dependence allows that simpler models can achieve good results maintaining nice interpretations of the model. In particular, the simple model (b) that only uses the easy-to-measure variate  $\text{Temp}_{n,s}(t)$ , shows that influenza may increase due to a cold wave with daily temperatures around  $7^\circ\text{C}$  for two weeks which is consistent with much of the literature on influenza. Also, the models show that the estimated temporal dependence of the influenza virus is strong and stable over time.

In our examples, we estimated the error structure with simple  $\text{AR}(p)$  models (mostly  $\text{AR}(1)$  or  $\text{AR}(2)$ ) obtaining a good fit for time dependence. We also tried other ARMA models and obtained similar results. Our method can additionally be used to explore more complex



dependence structures like heterogeneous covariances by counties or even spatio-temporal modelling. The iGLS procedure allows for more simplicity and flexibility in the estimation of the dependence structure at the cost of a light heavier computational work. Furthermore, in particular in the example provided, the iGLS allows us to specify a general dependence structure that can be adapted for every county rather than considering the same model for all counties or designing, by hand, the best structure for each county.

## Supporting information

**S1 Appendix. Complete description of functions, libraries, source data and code used along the paper.**

(PDF)

**S2 Appendix. Extended simulation.** Simulations results for model (b).

(PDF)

**S1 File. Supplemental code and data zip file.** File containing the code and dataset used along the paper.

(ZIP)

## Acknowledgments

The authors thanks the healthcare provider: the Service Epidemiology of the Dirección Xeral de Saúde Pública (SERGAS) from the Consellería de Sanidade (Xunta de Galicia).

## Author Contributions

**Conceptualization:** Manuel Oviedo de la Fuente, Manuel Febrero-Bande, María Pilar Muñoz, Àngela Domínguez.

**Data curation:** Manuel Oviedo de la Fuente.

**Formal analysis:** Manuel Oviedo de la Fuente, Manuel Febrero-Bande, María Pilar Muñoz.

**Funding acquisition:** Manuel Febrero-Bande.

**Methodology:** Manuel Oviedo de la Fuente, Manuel Febrero-Bande, María Pilar Muñoz, Àngela Domínguez.

**Software:** Manuel Oviedo de la Fuente, Manuel Febrero-Bande.

**Supervision:** Manuel Oviedo de la Fuente, Manuel Febrero-Bande, María Pilar Muñoz, Àngela Domínguez.

**Visualization:** Manuel Oviedo de la Fuente.

**Writing – original draft:** Manuel Oviedo de la Fuente, Manuel Febrero-Bande, María Pilar Muñoz, Àngela Domínguez.

## References

1. Watson JM, Pebody RG. Influenza surveillance and pandemic requirements. *Pandemic influenza*. 2012; p. 9–16. <https://doi.org/10.1079/9781845938567.0009>
2. Van-Tam J, Sellwood C. In: *Epidemiology and clinical features of interpandemic influenza*. CABI; 2012. p. 1–8.
3. Paget J, Marquet R, Meijer A, van der Velden K. Influenza activity in Europe during eight seasons (1999–2007): an evaluation of the indicators used to measure activity and an assessment of the timing,

- p>length and course of peak activity (spread) across Europe.
- BMC Infect Dis.*
- 2007; 7(1):1.
- <https://doi.org/10.1186/1471-2334-7-141>
4. Schaffer F, Soergel M, Straube D. Survival of airborne influenza virus: effects of propagating host, relative humidity, and composition of spray fluids. *Arch Virol.* 1976; 51(4):263–273. <https://doi.org/10.1007/BF01317930> PMID: 987765
  5. Monto AS, Pichichero ME, Blanckenberg SJ, Ruuskanen O, Cooper C, Fleming DM, et al. Zanamivir prophylaxis: An effective strategy for the prevention of influenza types A and B within households. *J Infect Dis.* 2002; 186(11):1582–1588. <https://doi.org/10.1086/345722> PMID: 12447733
  6. Choi K, Thacker SB. An evaluation of influenza mortality surveillance, 1962–1979 I. Time series forecasts of expected pneumonia and influenza deaths. *Am J Epidemiol.* 1981; 113(3):215–226. <https://doi.org/10.1093/oxfordjournals.aje.a113091> PMID: 6258426
  7. Dushoff J, Plotkin JB, Viboud C, Earn DJ, Simonsen L. Mortality due to influenza in the United States—an annualized regression approach using multiple-cause mortality data. *Am J Epidemiol.* 2006; 163(2):181–187. <https://doi.org/10.1093/aje/kwj024> PMID: 16319291
  8. Höhle M, Paul M. Count data regression charts for the monitoring of surveillance time series. *Comput Stat Data Anal.* 2008; 52(9):4357–4368. <https://doi.org/10.1016/j.csda.2008.02.015>
  9. Conesa D, Martínez-Beneito M, Amorós R, López-Quílez A. Bayesian hierarchical Poisson models with a hidden Markov structure for the detection of influenza epidemic outbreaks. *Stat Methods Med Res.* 2015; 24(2):206–223. <https://doi.org/10.1177/096228021414853> PMID: 21873301
  10. Ugarte M, Goicoa T, Militino A. Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics.* 2010; 21(3–4):270–289. <https://doi.org/10.1002/env.1011>
  11. Paul M, Held L. Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Stat Med.* 2011; 30(10):1118–1136. PMID: 21484849
  12. Cuevas A, Febrero M, Fraiman R. An ANOVA test for functional data. *Comput Stat Data Anal.* 2004; 47(1):111–122. <https://doi.org/10.1016/j.csda.2003.10.021>
  13. Ferraty F, Vieu P. *Nonparametric functional data analysis: theory and practice.* Springer; 2006.
  14. Chiou JM, Müller HG, Wang JL. Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2003; 65(2):405–423. <https://doi.org/10.1111/1467-9868.00393>
  15. Sørensen H, Goldsmith J, Sangalli LM. An introduction with medical applications to functional data analysis. *Stat Med.* 2013; 32(30):5222–5240. <https://doi.org/10.1002/sim.5989> PMID: 24114808
  16. Ramsay JO, Silverman BW. *Functional Data Analysis.* Springer; 2005.
  17. Cardot H, Ferraty F, Sarda P. Spline estimators for the functional linear model. *Stat Sinica.* 2003; 13(3):571–592.
  18. Antoniadis A, Sapatinas T. Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. *J Multivariate Anal.* 2003; 87(1):133–158. [https://doi.org/10.1016/S0047-259X\(03\)00028-9](https://doi.org/10.1016/S0047-259X(03)00028-9)
  19. Horváth L, Kokoszka P. *Inference for functional data with applications.* vol. 200. Springer; 2012.
  20. Cardot H, Ferraty F, Sarda P. Functional linear model. *Stat Probabil Lett.* 1999; 45(1):11–22. [https://doi.org/10.1016/S0167-7152\(99\)00036-X](https://doi.org/10.1016/S0167-7152(99)00036-X)
  21. Cardot H, Mas A, Sarda P. CLT in functional linear regression models. *Probab Theory Rel.* 2007; 138(3):325–361. <https://doi.org/10.1007/s00440-006-0025-2>
  22. Preda C, Saporta G. PLS regression on a stochastic process. *Comput Stat Data Anal.* 2005; 48(1):149–158. <https://doi.org/10.1016/j.csda.2003.10.003>
  23. Aneiros-Pérez G, Vieu P. Semi-functional partial linear regression. *Stat Probabil Lett.* 2006; 76(11):1102–1110. <https://doi.org/10.1016/j.spl.2005.12.007>
  24. Aneiros-Pérez G, Vieu P. Nonparametric time series prediction: A semi-functional partial linear modeling. *J Multivariate Anal.* 2008; 99(5):834–857. <https://doi.org/10.1016/j.jmva.2007.04.010>
  25. Müller HG, Yao F. Functional additive models. *J Am Stat Assoc.* 2008; 103(484):1534–1544. <https://doi.org/10.1198/016214508000000751>
  26. Ferraty F, Vieu P. Additive prediction and boosting for functional data. *Comput Stat Data Anal.* 2009; 53(4):1400–1413. <https://doi.org/10.1016/j.csda.2008.11.023>
  27. Febrero-Bande M, González-Manteiga W. Generalized additive models for functional data. *TEST.* 2013; 22(2):278–292. <https://doi.org/10.1007/s11749-012-0308-0>
  28. Chen D, Hall P, Müller HG. Single and multiple index functional regression models with nonparametric link. *Ann Stat.* 2011; 39(3):1720–1747. <https://doi.org/10.1214/11-AOS882>

29. Goia A. A functional linear model for time series prediction with exogenous variables. *Stat Probabil Lett*. 2012; 82(5):1005–1011. <https://doi.org/10.1016/j.spl.2012.02.009>
30. Ferraty F, Goia A, Salinelli E, Vieu P. Functional projection pursuit regression. *TEST*. 2013; 22(2):293–320. <https://doi.org/10.1007/s11749-012-0306-2>
31. Delicado P, Giraldo R, Comas C, Mateu J. Statistics for spatial functional data: some recent contributions. *Environmetrics*. 2010; 21(3–4):224–239. <https://doi.org/10.1002/env.1003>
32. Giraldo R, Delicado P, Mateu J. Ordinary kriging for function-valued spatial data. *Environ Ecol Stat*. 2011; 18(3):411–426. <https://doi.org/10.1007/s10651-010-0143-y>
33. Menafoglio A, Secchi P, Dalla Rosa M. A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electron J Stat*. 2013; 7:2209–2240. <https://doi.org/10.1214/13-EJS843>
34. Battey H, Sancetta A. Conditional estimation for dependent functional data. *J Multivariate Anal*. 2013; 120:1–17. <https://doi.org/10.1016/j.jmva.2013.04.009>
35. Besse PC, Cardot H, Stephenson DB. Autoregressive forecasting of some functional climatic variations. *Scand J Stat*. 2000; 27(4):673–687. <https://doi.org/10.1111/1467-9469.00215>
36. Damon J, Guillas S. Estimation and simulation of autoregressive hilbertian processes with exogenous variables. *Statistical Inference for Stochastic Processes*. 2005; 8(2):185–204. <https://doi.org/10.1007/s11203-004-1031-6>
37. Hörmann S, Kokoszka P. Weakly dependent functional data. *Ann Stat*. 2010; 38(3):1845–1884. <https://doi.org/10.1214/09-AOS768>
38. Kariya T, Kurata H. Generalized least squares. Wiley; 2004.
39. Febrero-Bande M, Oviedo de la Fuente M. Functional Regression Models with Temporal and/or Spatial Dependence. In *Contributions in infinite-dimensional statistics and related topics* pp:107–112. Società Editrice Esculapio, 2014.
40. Carmack PS, Spence JS, Schucany WR. Generalised correlated cross-validation. *J Nonparametr Stat*. 2012; 24(2):269–282. <https://doi.org/10.1080/10485252.2012.655733>
41. Pinheiro J, Bates D, DebRoy S, Sarkar D. R Core Team (2014) nlme: linear and nonlinear mixed effects models. R package version 3.1-117. Available at <https://cran.r-project.org/web/packages/nlme>. 2014;.
42. Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*. 1986; 73(1):43–56. <https://doi.org/10.1093/biomet/73.1.43>
43. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *Ann Statist*. 2007; 35(6):2769–2794. <https://doi.org/10.1214/009053607000000505>
44. Lyons R. Distance covariance in metric spaces. *Ann Probab*. 2013; 41(5):3284–3305. <https://doi.org/10.1214/12-AOP803>
45. Febrero-Bande M, Oviedo de la Fuente M. Statistical computing in functional data analysis: the R package fda.usc. *J Stat Softw*. 2012; 51(4):1–28. <https://doi.org/10.18637/jss.v051.i04>